

# Multi-agent Inverse Reinforcement Learning for Zero-sum Games

**Xiaomin Lin**  
**Peter A. Beling**  
**Randy Cogill**

*Department of Systems and Information Engineering*  
*University of Virginia*  
*Charlottesville, VA 22903 USA*

XL5DB@VIRGINIA.EDU  
 PB3A@VIRGINIA.EDU  
 RLC9S@VIRGINIA.EDU

## Abstract

In this paper we introduce a Bayesian framework for solving a class of problems termed Multi-agent Inverse Reinforcement Learning (MIRL). Compared to the well-known Inverse Reinforcement Learning (IRL) problem, MIRL is formalized in the context of a stochastic game rather than a Markov decision process (MDP). Games bring two primary challenges: First, the concept of optimality, central to MDPs, loses its meaning and must be replaced with a more general solution concept, such as the Nash equilibrium. Second, the non-uniqueness of equilibria means that in MIRL, in addition to multiple reasonable solutions for a given inversion model, there may be multiple inversion models that are all equally sensible approaches to solving the problem. We establish a theoretical foundation for competitive two-agent MIRL problems and propose a Bayesian optimization algorithm to solve the problem. We focus on the case of two-person zero-sum stochastic games, developing a generative model for the likelihood of unknown rewards of agents given observed game play assuming that the two agents follow a minimax bipolicy. As a numerical illustration, we apply our method in the context of an abstract soccer game. For the soccer game, we investigate relationships between the extent of prior information and the quality of learned rewards. Results suggest that covariance structure is more important than mean value in reward priors.

## 1. Introduction

Inverse Reinforcement Learning (IRL) problem, as characterized in (Russell, 1998), aims to optimally recover reward functions, given the measurements of an agent's behavior over time, as well as a model of the environment. IRL has been the subject of extensive studies and has been applied to a number of problems, most related to the problem of learning from demonstrations. Apprenticeship learning algorithms based on IRL, which leverage expert demonstrations to efficiently learn good controllers for tasks being demonstrated by an expert, have been applied to automatic control of helicopter flight (Abbeel & Ng, 2004) and modeling of driver route preferences (Ziebart, Maas, Bagnell, & Dey, 2008). In the field of computer graphics, IRL also has been used to learn behavior styles for the motion controller of an animation system (Lee & Popovic, 2010). In (Baker, Saxe, & Tenenbaum, 2009), IRL was viewed from the perspective of human decision making as a method for modeling human action understanding, and the results of psychophysical experiments using

animated stimuli of agents moving in simple masses provide quantitative evidence that the inverse planning models can predict human goal inferences.

Although a variety of approaches (Ng & Russell, 2000; Qiao & Beling, 2011; Levine, Popović, & Koltun, 2011; Krishnamurthy & Todorov, 2010; Ramachandran & Amir, 2007) have been proposed for solving the IRL problem, almost all of them are based on the assumption that no other adaptive agents exist in the environment. Attempting to jointly consider the decision making processes of interacting rational agents can significantly complicate models, but leads to more insightful models of multi-agent systems. This is the motivation for Littman to propose the *Multi-agent Reinforcement Learning* (MRL) problem (Littman, 1994). Conceptually, RL is an simplified or approximate version of MRL in the sense that the former treats other agents in the system as part of the environment, ignoring the difference between responsive agents and passive environment. Littman makes use of a Markov or stochastic game (Owen, 1968), which is an extension of game theory to *Markov Decision Process* (MDP)-like environments. However, only the special case of *two-player zero-sum games*, in which one agent’s gain is always the other’s loss, is considered. Hu and Wellman (Hu & Wellman, 1998) extend Littman’s work, proposing a *two-player general-sum* stochastic game framework for the MRL problem. They point out that the concept of optimality loses its meaning in MRL problems since any agent’s payoff depends on others’ choices of actions, and as a result adopt as a solution concept the *Nash equilibrium*, in which each agent’s choice is the best response to other agents’ choices.

Later MRL work has focused on the development of solution concepts and methods. For example, in (Abdallah & Lesser, 2008) a weak condition where an agent can neither observe other agents’ actions or rewards, nor knows the underlying game or the corresponding Nash equilibrium a priori is considered and a new MRL algorithm called the Weighted Policy Learner (WPL) is proposed. Multi-agent learning in complex large distributed systems is also touched in (Kash, Friedman, & Halpern, 2011), where it is noted that, although sophisticated multi-agent learning algorithms generally do not scale, it is possible to find restricted classes of games where simple efficient algorithms converge. Solution concepts for distributed, multi-agent planning problems that involve coordination games under weak information exchange models have been considered in (Patek, Beling, & Zhao, 2007; Zhao, Patek, & Beling, 2008).

Inverse learning problems for MRL, which we term MIRL, include the problem of estimating the game payoffs being played, given only observations of the actions taken by the players. Compared to the well-known IRL problem, MIRL is more challenging in that it is formalized in the context of a stochastic game rather than a MDP. Games bring two primary challenges. First, as Hu and Wellman (Hu & Wellman, 1998) note, the concept of optimality, central to MDPs, must be replaced with an equilibrium solution concept, such as the Nash equilibrium. Second, the non-uniqueness of equilibria means that in MIRL, in addition to multiple reasonable solutions for a given inversion model, there may be multiple inversion models that are all equally sensible approaches to solving the problem.

Several recent papers have studied problems that may appear quite similar to the MIRL problem discussed here. For example, (Natarajan, Kunapuli, Judah, Tadepalli, Kersting, & Shavlik, 2010) presents an inverse reinforcement learning model for multiple agents. However, that paper does not consider competing agents or game-theoretic models, a key characteristic of our work. The recent paper (Waugh, Ziebart, & Bagnell, 2011) does con-

sider a form of the inverse equilibrium problem. However, that paper considers simultaneous one-stage games, rather than the sequential stochastic games we consider here. The formulation of the MIRL problem and a Bayesian framework for its solution was first given by the authors in (Beling, Cogill, & Lin., 2013). This paper extends that work through inclusion of additional technical details and by reporting on a broader and deeper set of numerical experiments and simulations.

We model the MIRL problem in the framework of stochastic games. Unlike Markov decision processes, which generally have deterministic optimal policies (Bertsekas, 2005; Filar & Vrieze, 1996), equilibrium strategies in stochastic games may be mixed strategies. An extensive literature exists on stochastic games (see, e.g., (Bewley & Kohlberg, 1976; Federgruen, 1980; Filar & Vrieze, 1996; Kushner & Chamberlain, 1969; Maitra & Parthasarathy, 1970; Mertens & Neyman, 1980; Monash, 1979; Raghavan & Filar, 1991; Rao, Chandrasekaran, & Nair, 1973; Rogers, 1969; Shapley, 1953; Sobel, 1971; Vrieze, 1987)). Though stochastic games may be defined quite generally, we restrict attention to finite Markov stochastic games, as defined in Section 2.

In many stochastic games, the sum of the two agents’ rewards can be arbitrary. An often studied special case is a class of problems where the agents’ rewards sum to zero. These two cases correspond to two major topics in game theory, two-person general-sum games and two-person zero-sum games (Ferguson, 2008; Owen, 1968). In this paper, we restrict attention to the latter class of problems, proposing a mathematical formulation of the two-person zero-sum MIRL problem framed in the context of strictly competitive agents.

The primary contribution of this paper is a Bayesian optimization framework for posing and solving stochastic game-based, two-person, zero-sum MIRL problems. The Bayesian framework is well-defined for any prior. When restricted to Gaussian priors, however, the framework yields a tractable, convex optimization problem for computing point estimates of rewards.

To illustrate the concepts and to provide a context for experimentation on model sensitivity, we apply our method to an abstract soccer game. For a broad range of basic dynamics within the soccer model, we investigate relationships between the extent of prior information and the quality of learned rewards. The quality of learned rewards is measured by distance metrics in reward and probability space and by the game playing success of agents that use the rewards as the basis for an equilibrium policy. The weakest priors result in learned rewards that would give an agent using them no chance of winning the game, while the strongest priors result in learned rewards essentially as good as ground truth. Additionally, results suggest that covariance structure is more important than mean value in reward priors.

The remainder of the paper is structured as follows: Section 2 introduces notation, terminology, definitions, and some basic properties needed for later work. Section 3 provides the main technical results, including a Bayesian framework for MIRL and formulation of a convex optimization problem for learning rewards. Section 4 introduces the soccer model and includes basic results on the quality of learned rewards. Section 5 provides evaluation of learned rewards in terms of game playing success in simulations of the soccer game. Section 6 offers concluding remarks and a discussion of future work.

## 2. Stochastic Games

A two-player *discounted stochastic game* is played as follows. The game begins in one of finitely many states. There is a reward function associated with each state for each player, and each player has perfect knowledge of the others rewards. Simultaneously, each player selects one of finitely many actions, and each player receives a reward that depends on the current state and sometimes, as well as the actions selected by both players. The game then makes a stochastic transition to a new state, where the transition is dependent on the starting state and the jointly selected actions. This process is repeated over an infinite time horizon, where geometrically discounted rewards are accrued additively.

Under these rules, we can specify an instance of a zero-sum stochastic game in terms of the state space  $\mathcal{S} = \{1, 2, \dots, N\}$ , the action spaces  $\mathcal{A}_1 = \mathcal{A}_2 = \{1, 2, \dots, M\}$ , reward functions  $r^k : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \mapsto \mathbb{R}$  for each player  $k \in \{1, 2\}$ , transition probabilities  $p(s'|s, a^1, a^2)$ , and a discount factor  $\gamma \in [0, 1)$ .

A solution to a stochastic game is a *bipolicy*, which provides the rules that each player follows when selecting actions at each state. Without loss of generality, a bipolicy can be specified by a collection of conditional probability mass functions  $\pi^1$  and  $\pi^2$ , where player  $k$  selects action  $a^k$  in state  $s$  with probability  $\pi^k(a^k|s)$ . Each  $\pi^k(\cdot|s)$  is referred to as the *strategy* played by player  $k$  in state  $s$ .

Given that each player can select from among  $M$  actions, the strategy followed by player  $k$  in state  $s$  can be represented by the  $M \times 1$  vector  $\pi^k(s)$ . The bistrategy for state  $s$  is the set of two column vectors that denote the strategies employed by player 1 and player 2 in state  $s$ ,

$$\pi(s) = \{\pi^1(s), \pi^2(s)\}.$$

In this notation, the bipolicy is defined as the set of all bistrategies over all states,

$$\pi = \{\pi(1), \pi(1), \dots, \pi(N)\}.$$

### 2.1 Zero-sum Case

A *two-player zero-sum discounted stochastic game* is a special case of the game defined above, in which under the same state  $s$  and action pairs  $(a^1, a^2)$ ,  $r^1(s, a^1, a^2) = -r^2(s, a^1, a^2)$ . Due to the symmetry in rewards between the two players, we will often simply use  $r$  to denote  $r^1$ . In the remainder of this paper restrict attention to the zero-sum case.

We use  $\tilde{r}_\pi(s)$  to denote the single-stage *expected reward* of agent 1 under state  $s$  under bipolicy  $\pi$ , and  $\tilde{r}_\pi$  as a column vector with  $i$ th component  $\tilde{r}_\pi(s)$ .  $\tilde{r}_\pi(s)$  is defined as

$$\tilde{r}_\pi(s) = \sum_{a^1, a^2} \pi^1(a^1|s) \pi^2(a^2|s) r(s, a^1, a^2) = [\pi^1(s)]^T r(s) \pi^2(s). \quad (1)$$

We can express this relationship in matrix notation as

$$\tilde{r}_\pi = B_\pi r, \quad (2)$$

where  $B_\pi$  is a  $N \times NM^2$  matrix constructed from bipolicy  $\pi$ , whose  $k$ th row is:

$$[\Phi_{1,1}^\pi(k), \Phi_{1,2}^\pi(k), \dots, \Phi_{M,M}^\pi(k)],$$



where

$$\Phi_{i,j}^{\pi}(k) = \underbrace{0, \dots, 0}_{k-1}, \phi_{i,j}^{\pi}(k), \underbrace{0, \dots, 0}_{N-k},$$

and

$$\phi_{i,j}^{\pi}(k) = \pi^1(i|k) \pi^2(j|k).$$

The bipolicy-dependent, discounted expected sum of rewards of player 1 as a function of the initial state, which is known as the value function, can be formulated as:

$$V_{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t E(r_{\pi}^t(s_t) | s_0 = s), \quad (3)$$

where  $s_t$  denotes the state of the game at stage  $t$ .  $V_{\pi}$  denotes the column vector with  $i$ th component  $V_{\pi}(i)$ .

In addition, we define player 1's Q-function of state  $s$  and action pair  $(a^1, a^2)$ , under bipolicy  $\pi$ , as

$$Q_{\pi}(s, a^1, a^2) = r(s, a^1, a^2) + \gamma \sum_{s'} p(s'|s, a^1, a^2) V_{\pi}(s'). \quad (4)$$

Over all states and actions, we can write equation (4) in matrix notation as

$$Q_{\pi} = r + \gamma P V_{\pi}, \quad (5)$$

where  $P$  is a  $NM^2 \times N$  matrix with  $p(s'|s, a^1, a^2)$  as its elements.

Let  $G_{\pi}$  denote transition matrix under bipolicy  $\pi$ . Specifically,  $G_{\pi}$  is the  $N \times N$  matrix with elements

$$g_{\pi}(s'|s) = \sum_{a^1, a^2} \pi^1(a^1|s) \pi^2(a^2|s) p(s'|s, a^1, a^2). \quad (6)$$

Note that

$$\begin{aligned} V_{\pi}(s_0) &= \tilde{r}_{\pi}(s_0) + \gamma \left\{ \sum_{t=1}^{\infty} \gamma^{t-1} E(r_{\pi}^t(s_t) | s_0 = s) \right\} \\ &= \tilde{r}_{\pi}(s_0) + \gamma \sum_{s'} g_{\pi}(s'|s) V_{\pi}(s'). \end{aligned} \quad (7)$$

This equation can be written in matrix notation as

$$V_{\pi} = \tilde{r}_{\pi} + \gamma G_{\pi} V_{\pi}. \quad (8)$$

Thus

$$V_{\pi} = (I - \gamma G_{\pi})^{-1} B_{\pi} r, \quad (9)$$

where  $(I - \gamma G_{\pi})$  is always invertible for  $\gamma \in [0, 1)$  since  $G_{\pi}$  is a transition matrix. The value function  $V_{\pi}(s)$  can be expressed in terms of the Q-function as

$$V_{\pi}(s) = [\pi^1(s)]^T Q_{\pi}(s) \pi^2(s). \quad (10)$$

where  $Q_{\pi}(s)$  is a  $M \times M$  matrix for agent 1, whose  $(i, j)$  element is given by  $Q_{\pi}(s, i, j)$ . Note that while  $Q_{\pi}(s)$  is a matrix, the  $Q_{\pi}$  introduced in (5) is an  $NM^2 \times 1$  vector. We will

use this relationship between the  $Q$ -function and the value function to define a *minimax bipolicy* for a stochastic game.

We will assume that rational agents playing a two-player zero-sum stochastic game seek a minimax bipolicy. A minimax bipolicy is an equilibrium, in that it has the property that neither player can change the game value in their favor given that the other player holds their policy fixed. To give a precise definition of a minimax bipolicy, we will start by reviewing the notion of a minimax bistrategy for a static game (Neumann & Morgenstern, 1944).

First consider a static (single-stage) zero-sum game, where two players simultaneously choose an action and both players receive a reward determined by the joint choice of actions. The minimax theorem states that for every two-person zero-sum game with finitely many actions, there exists a value  $V$  and a mixed strategy for each player such that

- Given player 2's strategy, the best expected reward possible for player 1 is  $V$ .
- Given player 1's strategy, the best expected reward possible for player 2 is  $-V$ .

As before, the strategies played by both players in a certain state  $s$  can be expressed in terms of probability mass functions  $\pi^1$  and  $\pi^2$ . Expressing the reward received by player 1 as an  $M \times M$  matrix  $Q$ , the value of the game for player 1 under a minimax bistrategy is given by

$$\text{value}(Q) = \max_{\pi^1} \left\{ \min_{\pi^2} \left\{ [\pi^1]^T Q \pi^2 \right\} \right\}.$$

A pair  $\pi^1$  and  $\pi^2$  that achieves this value is called a *minimax bistrategy*. For zero-sum games, a minimax bistrategy is also a Nash equilibrium.

The concept of a minimax bistrategy can be extended to two-player discounted stochastic games via the following theorem (Shapley, 1953).

**Theorem 2.1** (Shapley's Theorem). *There exists a bipolicy  $\pi$  such that*

$$V_\pi(s) = \text{value}(Q_\pi(s)) \tag{11}$$

for all  $s \in \mathcal{S}$ .

A bipolicy that satisfies Theorem 2.1 is called a *minimax bipolicy*. For a minimax bipolicy,  $V_\pi(s)$  gives the game value from each initial state  $s \in \mathcal{S}$ . In the remainder of this paper, we will assume that agents are observed playing a game according to a minimax bipolicy, and the known properties of their bipolicy will be used to infer the reward structure of the game.

### 3. Bayesian MIRL

We will formulate two-agent MIRL problems in a Bayesian optimization setting. Bayesian methods have been widely adopted for IRL problems (Baker et al., 2009; Choi & Kim, 2011; Dimitrakakis & Rothkopf, 2011; Engel, Mannor, & Meir, 2005; Michini & How, 2012; Qiao & Beling, 2011; Ramachandran & Amir, 2007). In a Bayesian setting, we assign a prior distribution to the reward functions. This prior distribution encodes the learners initial

belief about the reward functions before any observations are made. For our problem, we assume that the complete bipolicy for the agents is observable. Given an observed bipolicy, we can generate a point estimate of the reward function from the posterior distribution over reward functions. To construct this point estimate, we must know the likelihood of observing each bipolicy for each given reward function. So, our efforts are focused on determining the appropriate likelihood function for the MIRL problem, and the development of optimization models that can be used to generate point estimates of the reward function.

We will now formally develop the Bayesian model for the MIRL problem. Let  $f(r)$  denote the prior distribution on the reward of agent 1 (recalling that we denote  $r = r^1$  and  $r^1 = -r^2$  for zero-sum games). We will discuss the selection of prior distributions further in Section 3.1. Also, let  $p(\pi|r)$  denote the likelihood of observing bipolicy  $\pi$  when the true reward is  $r$ . Hence now our objective is to maximize  $f(r|\pi)$ , the posterior of rewards given an observing bipolicy. To model the likelihood  $p$ , we make the following assumptions regarding the agents selection of a bipolicy:

1. The two agents select a minimax bipolicy with respect to the reward function  $r$ .
2. If the minimax bipolicy for a given  $r$  is not unique, then one is selected uniformly at random from among all minimax bipolicies.

If the minimax bipolicy is unique for a given  $r$ , the likelihood is a probability mass function given by

$$p(\pi|r) = \begin{cases} 1, & \text{if } \pi \text{ is the unique minimax bipolicy for } r \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

If the minimax bipolicy is not unique, then there are an uncountable number of minimax bipolicies. In this case, the likelihood is a uniform distribution given by

$$p(\pi|r) = \begin{cases} \frac{1}{K(r)}, & \text{if } \pi \text{ is a minimax bipolicy for } r \\ 0, & \text{otherwise,} \end{cases}$$

where  $K(r)$  is a normalizing constant.

The posterior distribution of rewards for a given observed bipolicy is now

$$f(r|\pi) \propto p(\pi|r) f(r).$$

So

$$f(r|\pi) \propto \begin{cases} \frac{f(r)}{K(r)}, & \text{if } \pi \text{ is a minimax bipolicy for } r \\ 0, & \text{otherwise,} \end{cases}$$

where we have used  $K(r) = 1$  for the case where the minimax bipolicy is unique.

A maximum a-posteriori (MAP) estimate of  $r$  can be generated by selecting an  $r$  that maximizes  $f(r|\pi)$ , or equivalently maximizes  $p(\pi|r) f(r)$ . The next sections provide the details of an optimization formulation of the MAP estimation approach to MIRL.

### 3.1 Prior Distributions on Rewards

In our Bayesian MIRL framework, we use prior distributions over reward functions to model our initial uncertainty in the reward. Although any prior may be used, in this paper we focus on Gaussian priors for rewards. Gaussians are a reasonable choice of prior since they provide a straightforward model for representing uncertainty around a nominal choice of reward function, and have the added benefit of leading to analytically tractable inference procedures.

Specifically, we model  $r \sim \mathcal{N}(\mu_r, \Sigma_r)$ , where  $\mu_r$  is the mean of  $r$  and  $\Sigma_r$  is the covariance matrix. The probability density function of  $r$  is

$$f(r) = \frac{1}{(2\pi)^{N/2} |\Sigma_r|^{1/2}} \exp\left(-\frac{1}{2} (r - \mu_r)^T \Sigma_r^{-1} (r - \mu_r)\right). \quad (13)$$

### 3.2 Characterization of the Likelihood Function

Recall that the likelihood  $p(\pi|r)$  is nonzero if and only if  $\pi$  is a minimax bipolicy with respect to the reward function  $r$ . In this paper, we will assume that there is a unique minimax bipolicy for all reward functions of interest. When this is the case,  $p(\pi|r) = 1$  when  $\pi$  is the unique minimax bipolicy for the reward function  $r$  and  $p(\pi|r) = 0$  otherwise.

### 3.3 The Optimization Model

In this section, we will present the details of an optimization formulation of the problem of computing a MAP estimate of  $r$  given a bipolicy  $\pi$ . Computing a MAP-optimal reward function reduces to computing an  $r$  that maximizes  $p(\pi|r) f(r)$ , where  $f(r)$  is a prior distribution over rewards. So, in the case where the given minimax bipolicy is unique with respect to the desired reward, a MAP-optimal reward can be computed by solving the optimization problem

$$\begin{aligned} &\text{maximize: } f(r) \\ &\text{subject to: } p(\pi|r) = 1 \end{aligned} \quad (14)$$

The remainder of this section will be devoted to developing a tractable characterization of the set of feasible  $r$ . Consider, as a first step, the class of static, single-stage, zero-sum games. In these games, minimax strategies satisfy the conditions of the following theorem (Neumann & Morgenstern, 1944; Ferguson, 2008):

**Theorem 3.1** (Minimax Theorem). *Consider a two-person zero-sum game with  $M \times M$  matrix  $A$ . There exists a value  $V$ , a mixed strategy  $p$  for player 1, and a mixed strategy  $q$  for player 2 such that*

$$\begin{aligned} A^T p &\geq V \mathbf{1}_M \\ A q &\leq V \mathbf{1}_M \end{aligned} \quad (15)$$

Moreover,  $p$  and  $q$  are an equilibrium bistrategy and  $V$  is the game value if and only if (15) holds.

This theorem has direct implications for inverse learning problems. Consider a static game as a special case of the MIRL problem, where the goal is to recover a  $A$  such that the given bistrategy  $(p, q)$  is a minimax bistrategy. In this case, (15) must be satisfied

by  $A$  and  $V$  if  $A$  represents the reward matrix and  $V$  is the game value, with  $(p, q)$  as a minimax bistrategy. Hence, the linear constraints (15) give a characterization of the desired constraint set for a two-person zero-sum static game.

We will now extend this approach to a multi-stage stochastic game. Combining Theorem 2.1 with Theorem 3.1, a bipolicy  $\pi$  is a minimax bipolicy if and only if

$$\begin{aligned} [Q_\pi(s)]^T \pi^1(s) &\geq V_\pi(s) \mathbf{1}_M \\ Q_\pi(s) \pi^2(s) &\leq V_\pi(s) \mathbf{1}_M \end{aligned} \quad (16)$$

for all  $s \in \mathcal{S}$ . The linear inequalities (16) provide conditions that must hold for the  $Q$ -function and value function of a stochastic game if  $\pi$  is a minimax bipolicy.

Since our ultimate goal is to estimate the reward function of a stochastic game, we must introduce additional constraints relating the  $Q$ -function and value function to rewards. From (5) and (9), recall that

$$\begin{aligned} Q_\pi &= r + \gamma P V_\pi \\ V_\pi &= (I - \gamma G_\pi)^{-1} B_\pi r \end{aligned} \quad (17)$$

The following proposition makes use of the linear inequalities (16) and equalities (17) to provide a precise relationship between the given bipolicy  $\pi$  and the rewards  $r$  if  $\pi$  is a minimax bipolicy.

**Proposition 3.2.** *In a two-person zero-sum stochastic game,  $\pi$  is a minimax bipolicy if and only if player 1's reward vector  $r$  satisfies*

$$\begin{aligned} (B_{\pi^2|a^1=i} - B_\pi) D_\pi r &\leq 0 \\ (B_{\pi^1|a^2=j} - B_\pi) D_\pi r &\geq 0 \end{aligned} \quad (18)$$

for all  $i \in \mathcal{A}_1$  and  $j \in \mathcal{A}_2$ , where  $B_\pi$  is defined in (2).  $B_{\pi^1|a^2=j}$  and  $D_\pi$  are defined in (20) and (22) in Appendix A, respectively.

**Proof.** From (1), (2) and (10), we can deduce that

$$V_\pi = B_\pi Q_\pi. \quad (19)$$

Let  $B_{\pi^1|a^2=j}$  denote the  $B_\pi$  obtained when  $\pi^1$  is used as player 1's policy, and player 2 selects action  $a^2 = j$  in all states. In this notation, the inequalities (15) can be expressed as

$$\begin{aligned} B_{\pi^1|a^2=j} Q_\pi &\geq B_\pi Q_\pi, \forall j \in \mathcal{A}_2 \\ B_{\pi^1|a^1=i} Q_\pi &\leq B_\pi Q_\pi, \forall i \in \mathcal{A}_1 \end{aligned} \quad (20)$$

Substituting the expression for  $V_\pi$  into the expression for  $Q_\pi$  in (17), we obtain

$$Q_\pi = r + \gamma P (I - \gamma G_\pi)^{-1} B_\pi r = \left( I + \gamma P (I - \gamma G_\pi)^{-1} B_\pi \right) r. \quad (21)$$

Finally, letting

$$D_\pi = \left( I + \gamma P (I - \gamma G_\pi)^{-1} B_\pi \right), \quad (22)$$

the inequalities (20) can be expressed as

$$\begin{aligned} (B_{\pi^1|a^2=j} - B_\pi) D_\pi r &\geq 0, \forall j \in \mathcal{A}^2 \\ (B_{\pi^2|a^1=i} - B_\pi) D_\pi r &\leq 0, \forall i \in \mathcal{A}^1 \end{aligned} \quad (23)$$

We now formulate a convex quadratic program equivalent to (14). Recall that we use a Gaussian prior in this paper, so the objective function in (14) is log-concave. To obtain an equivalent convex optimization problem, we will instead minimize  $-\ln(f(r))$ . Combining (18) with the negative log-prior objective, the optimization problem (14) can be solved as the equivalent convex quadratic program

$$\begin{aligned} \text{minimize: } & \frac{1}{2} (r - \mu_r)^T \Sigma_r^{-1} (r - \mu_r) \\ \text{subject to: } & (B_{\pi^2|a^1=i} - B_\pi) D_\pi r \leq 0 \\ & (B_{\pi^2|a^1=j} - B_\pi) D_\pi r \geq 0 \end{aligned} \quad (24)$$

for all  $i \in \mathcal{A}_1$  and  $j \in \mathcal{A}_2$ .

The optimization problem (24) is specific to two-person zero-sum MRL problems, which is a class of problems in which the reward value does depend on actions. The following proposition provides a variant of Proposition 3.2 that applies when reward does not depend on actions.

**Proposition 3.3.** *In a two-person zero-sum stochastic game, assume that the reward is only state-dependent, then  $\pi$  is a minimax bipolicy if and only if player 1's reward vector  $r$  satisfies*

$$\begin{aligned} (G_\pi - G_{\pi^2|a^1=i}) (I - \gamma G_\pi)^{-1} r &\geq 0 \\ (G_\pi - G_{\pi^1|a^2=j}) (I - \gamma G_\pi)^{-1} r &\leq 0 \end{aligned} \quad (25)$$

for all  $i \in \mathcal{A}_1$  and  $j \in \mathcal{A}_2$ , where  $G_\pi$  and  $G_{\pi^2|a^1=i}$  are defined in (6) and (31), respectively.

**Proof.** Recall that

$$\begin{aligned} [Q_\pi(s)]^T \pi^1(s) &\geq V_\pi(s) 1_M \\ Q_\pi(s) \pi^2(s) &\leq V_\pi(s) 1_M \end{aligned} \quad (26)$$

where

$$Q_\pi(s, a^1, a^2) = r(s) + \gamma \sum_{s'} p(s'|s, a^1, a^2) V_\pi(s') \quad (27)$$

$$V_\pi = r + \gamma G_\pi V_\pi. \quad (28)$$

Note that the  $Q$ -function (27) and the value function (28) are expressions for the case when rewards are independent of actions. Because of symmetry, we only consider the second inequality of (26), which can be expended as

$$\begin{bmatrix} Q_\pi(s, 1, 1), & \cdots, & Q_\pi(s, 1, M) \\ Q_\pi(s, 2, 1), & \cdots, & Q_\pi(s, 2, M) \\ \vdots & \vdots & \vdots \\ Q_\pi(s, M, 1), & \cdots, & Q_\pi(s, M, M) \end{bmatrix} \begin{bmatrix} \pi^2(s, 1) \\ \pi^2(s, 2) \\ \vdots \\ \pi^2(s, M) \end{bmatrix} \leq V_\pi(s) 1_M \quad (29)$$

Substituting (27) into (29) we obtain

$$r(s) \mathbf{1}_M + \gamma \begin{bmatrix} \sum_{s', a^2} p(s'|s, 1, a^2) \pi^2(a^2|s) V_\pi(s') \\ \sum_{s', a^2} p(s'|s, 2, a^2) \pi^2(a^2|s) V_\pi(s') \\ \vdots \\ \sum_{s', a^2} p(s'|s, M, a^2) \pi^2(a^2|s) V_\pi(s') \end{bmatrix} \leq V_\pi(s) \mathbf{1}_M. \quad (30)$$

Let

$$G_{\pi^2|a^1=i}(s, s') = g_{\pi^2|a^1=i}(s'|s) = \sum_{a^2} \pi^2(a^2|s) p(s'|s, i, a^2), \quad (31)$$

$G_{\pi^2|a^1=i}$  can be regarded as a *stochastic state transition matrix* under the condition that player 1 complies with a fixed policy that she always takes action  $i$  in any state while the other agent sticks to her original policy  $\pi^2$ . (30) then can be simplified as

$$r(s) \mathbf{1}_M + \gamma \begin{bmatrix} \sum_{s'} g_{\pi^2|a^1=1}(s'|s) V_\pi(s') \\ \sum_{s'} g_{\pi^2|a^1=2}(s'|s) V_\pi(s') \\ \vdots \\ \sum_{s'} g_{\pi^2|a^1=M}(s'|s) V_\pi(s') \end{bmatrix} \leq V_\pi(s) \mathbf{1}_M, \quad (32)$$

which can be rewritten as

$$r(s) + \sum_{s'} g_{\pi^2|a^1=i}(s'|s) V_\pi(s') \leq V_\pi(s), \forall i \in \mathcal{A}^1. \quad (33)$$

We can express the above inequality in matrix notation, as

$$r + \gamma G_{\pi^2|a^1=i} V_\pi \leq V_\pi. \quad (34)$$

Finally we can deduce the following inequality from (28) and (34)

$$(G_\pi - G_{\pi^2|a^1=i})(I - \gamma G_\pi)^{-1} r \geq 0, \forall i \in \mathcal{A}^1.$$

Symmetrically, we also have

$$(G_\pi - G_{\pi^1|a^2=j})(I - \gamma G_\pi)^{-1} r \leq 0, \forall j \in \mathcal{A}^2.$$

As a consequence, we can develop a new version of the optimization formulation by replacing the linear constraints in (24) with (25), as follows:

$$\begin{aligned} & \text{minimize:} \quad \frac{1}{2} (r - \mu_r)^T \Sigma_r^{-1} (r - \mu_r) \\ & \text{subject to:} \quad (G_\pi - G_{\pi^2|a^1=i})(I - \gamma G_\pi)^{-1} r \geq 0 \\ & \quad \quad \quad (G_\pi - G_{\pi^1|a^2=j})(I - \gamma G_\pi)^{-1} r \leq 0 \end{aligned}$$

for all  $i \in \mathcal{A}_1$  and  $j \in \mathcal{A}_2$ .

## 4. Numerical Example

In this section, we demonstrate the Bayesian MIRL method developed in the previous sections on a two-player stochastic game modeled on soccer. Though styled after soccer abstractions in (Littman, 1994; Beling et al., 2013), the game considered here is richer in that it models an action *shoot*, which is a direct attempt to score through a ball kick.

### 4.1 Game and Model

The game is played on a  $4 \times 5$  grid as depicted in Figure 1. We use A and B to denote two players, and the circle in the figures to represent the ball. Each player can either stay unmoved or move to one of its neighborhood squares by taking one of 5 actions in each turn: *N* (north), *S* (south), *E* (east), *W* (west), and *stand*. If both players land on the same square in the same time period, the ball is exchanged between the two players with probability  $\beta$ . In addition, the player who has the ball can elect to *shoot*, which is to kick the ball toward their opponent’s goal, with a *probability of succesful shot* (PSS) distribution shown in Table 1. Note that the action *shoot* may be taken from any field position, and the PSS is independent of opponent position. Both players act simultaneously in each time period. There are in total 800 states in this model, corresponding to the positions of the players and ball possession. Each players aims to maximize expected goals scored, subject to discount factor of  $\gamma = 0.9$ .

We aim to recover the rewards for player A (and hence for B) under the following assumptions:

1. This is a zero-sum game.
2. The ball exchange rate is known.
3. The bipolicy followed by the players is known and is minimax.

Without any knowledge of the point structure of the game, we will infer which squares each player must reach in order to score a point (the goal squares), as well as the PSS of each player.

Both players follow policies that attempt to dribble or shoot the ball into specific squares representing their opponent’s goal. Player A attempts to score by reaching or shooting the ball into squares 6 or 11, and player B attempts to score by reaching or shooting the ball into squares 10 or 15. Once a point is scored, the players take the positions shown in Figure 1 and ball possession is assigned randomly. By observing this policy alone, the MIRL algorithm should infer a reward function for player A that assigns positive reward when player A brings the ball to squares 6 or 11, negative reward when player B brings the ball to squares 10 or 15, and zero reward everywhere else. The PSS distribution of each player should also be recovered from the inferred reward.

For the soccer example, it is worth considering whether IRL could be used to learn rewards, and whether MIRL would offer any advantages in this regard. Figure 2 helps to answer these questions. A’s policy is not only decided by her own rewards, but also dependent on B’s responses to her policy, which is, essentially, controlled by B’s own rewards. And so is B’s policy. In brief, each player’s policy, is decided by its own rewards and those of its opponent. With conventional IRL the actions of the opponent could be modeled as



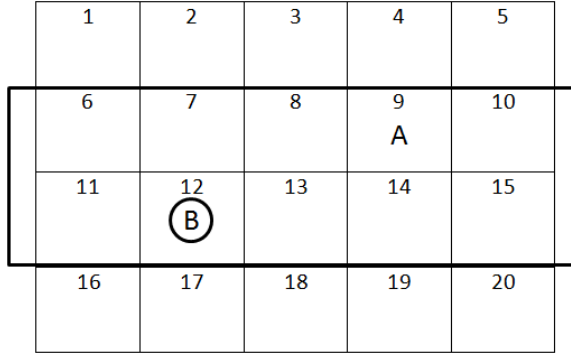


Figure 1: Soccer game: initial board

|   | PSS = 1 | PSS = 0.7    | PSS = 0.5    | PSS = 0.3    | PSS = 0.1     | PSS = 0 |
|---|---------|--------------|--------------|--------------|---------------|---------|
| A | 6, 11   | 1, 7, 12, 16 | 2, 8, 13, 17 | 3, 9, 14, 18 | 4, 10, 15, 19 | 5, 20   |
| B | 10, 15  | 5, 9, 14, 20 | 4, 8, 13, 19 | 3, 7, 12, 18 | 2, 6, 11, 17  | 1, 16   |

Table 1: Original PSS distribution of each player

part of the state transition probabilities, but would therefore need to be fixed with respect a given policy.

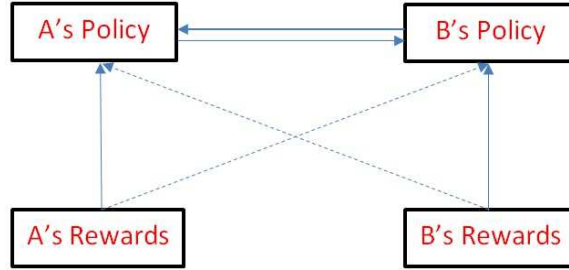


Figure 2: Soccer game internal relationship

## 4.2 Specification of Prior Information

Recall that the MRL optimization program requires the specification of two Gaussian prior parameters for A, the mean of the rewards vector  $\mu_r$  and the covariance matrix  $\Sigma_r$ . Below we define a concept of strength for prior information that can be expressed independently in the mean and covariance matrix. Later subsections focus on the impact of different priors on the quality of learned rewards.

### 4.2.1 MEAN OF THE PRIOR

We will use three types of mean reward vectors, namely *weak mean*, *median mean* and *strong mean*, respectively. Note that since this is a zero-sum game, the rewards assigned to B are the negatives of these rewards assigned to A.

- *Weak Mean*: we assign 0.8 point to player A in every state where A has possession of the ball and  $-0.8$  point in every state where player B has possession of the ball;
- *Median Mean*: guessing that A's goal might be among the rightmost squares, or squares 5, 10, 15 and 20, and symmetrically, B's goal might be among the leftmost squares, or squares 1, 6, 11 and 16, we assign 1 point to A whenever A has the ball and is in the four leftmost squares, and  $-1$  point to A whenever B has the ball and is in four rightmost squares. Also, when A has the ball and takes a shot, no matter where she is, we assign 0.5 point to A. Similarly, we assign  $-0.5$  point to A when B has the ball and takes a shot. Otherwise, no points will be assigned to A.
- *Strong Mean*: we have a foresight to predict where the goals are for both players, but cannot a good guess of their PSS distributions. So comparing to *median mean*, the only difference is that now the potential goal area includes only 2 squares (square 6 and 11 for A and square 10 and 15 for B), rather than 4 squares, for both players.

#### 4.2.2 COVARIANCE MATRIX

The covariance matrix of the reward vector encodes our belief of the structure of the prior. Based off of our knowledge of this soccer game, we can develop two types of covariance matrices.

- *Weak Covariance Matrix*: an identity matrix, indicating that the reward vector is assumed independently distributed. This is a universal covariance matrix suitable for those MRL problems in which we neither have knowledge of the structure of unknowns, nor want to make a guess.
- *Strong Covariance Matrix*: a more complex matrix encapsulating some internal information subject to our following beliefs.
  1. When A has the ball and takes a shot, the PSS depends only on A's position in the field; likewise for B.
  2. In any state, the reward for A for any non-*shoot* action is a state-dependent constant; likewise for B.

There are three types of relationships between any pair of rewards, perfect positive correlation, perfect negative correlation and mutual independence. In addition, we assume that the standard deviation of each random variable in the reward vector is the same. This assumption leads to a conclusion that normalized by a constant, the covariance matrix is equivalent to the correlation matrix of the unknown vector. Each entry in the correlation matrix is 0, 1 or -1, in accordance with the relationships embodied in the definitions of weak and strong covariance. Given the symmetry of the soccer problem, the covariance matrix  $\Sigma$  will be singular. This presents a problem because the MRL optimization procedure requires  $\Sigma^{-1}$  as input. We address this issue by working with the nonsingular matrix  $\hat{\Sigma} = \alpha I + \Sigma$ , where  $\alpha$  is a small positive scalar.

### 4.3 Results Evaluation Metric

To evaluate a recovered result, we simply compute its *Average Reward Distance* (ARD), which is the average *Euclidean distance* from the true rewards as follows.

$$\text{ARD} = \left\{ \frac{1}{2NM^2} \left[ (r_{\text{rec}}^1 - r^1)^T (r_{\text{rec}}^1 - r^1) + (r_{\text{rec}}^2 - r^2)^T (r_{\text{rec}}^2 - r^2) \right] \right\}^{1/2}, \quad (35)$$

where the  $NM^2 \times 1$  column vector  $r_{\text{rec}}^k$  and  $r^k$  denote the recovered and original reward of player  $k$ . Obviously, the smaller the ARD is, the more accurate the result is.

If only the players' PSS distributions are of interest, a similar version of the evaluation metric, termed *Average PSS Distance* (APD) can be defined as

$$\text{APD} = \left\{ \frac{1}{40} \left[ \sum_{i=1}^{20} (\theta_{\text{rec}}^1(i) - \theta_0^1(i))^2 + (\theta_{\text{rec}}^2(i) - \theta_0^2(i))^2 \right] \right\}^{1/2}, \quad (36)$$

where the  $20 \times 1$  column vector  $\theta_{\text{rec}}^k$  and  $\theta_0^k$  denote the recovered and original PSS of player  $k$ , respectively.

### 4.4 Results

Experiments were performed on 6 different priors formed by combining 3 different means and 2 different covariance matrices. An  $\alpha = 10^{-4}$  was used in the construction of the strong covariance matrices. In all cases, the bipolicy followed by the players (the observed input to MIRL) was computed iteratively from Shapley's Theorem, discussed in Section 2.1.

Results are shown in Figures 3-8. In each figure, the left subfigure shows the inferred rewards in blue and the benchmark rewards in red. The right subfigure shows the original PSS as a red stem and the inferred PSS as a blue stem. Table 2 sorts each experiment with a case number, maps each case to a figure and computes the corresponding APD.

|             | Weak Covariance          | Strong Covariance        |
|-------------|--------------------------|--------------------------|
| Weak Mean   | Case 1, Figure 3, 0.4535 | Case 2, Figure 4, 0.0671 |
| Median Mean | Case 3, Figure 5, 0.2169 | Case 4, Figure 6, 0.0387 |
| Strong Mean | Case 5, Figure 7, 0.2058 | Case 6, Figure 8, 0.0259 |

Table 2: Basic results summary ( $\alpha = 0.0001$ )

In Case 4, we are also interested in whether the MIRL algorithm can recover the actual goals for A and B. We calculate the average reward A receives when A is in square 1, 6, 11 and 16 and the average reward B receives when B is in square 5, 10, 15 and 20. The result is shown in Figure 9. The original rewards are shown as red stems and the inferred rewards blue ones.

It is also interesting to consider how the ball exchange rate  $\beta$  affects the PSS recovery result. We repeat Case 6 by changing  $\beta$  from 0 to 1, and calculate the APD of the inferred PSS distributions. The result is shown in Figure 10.

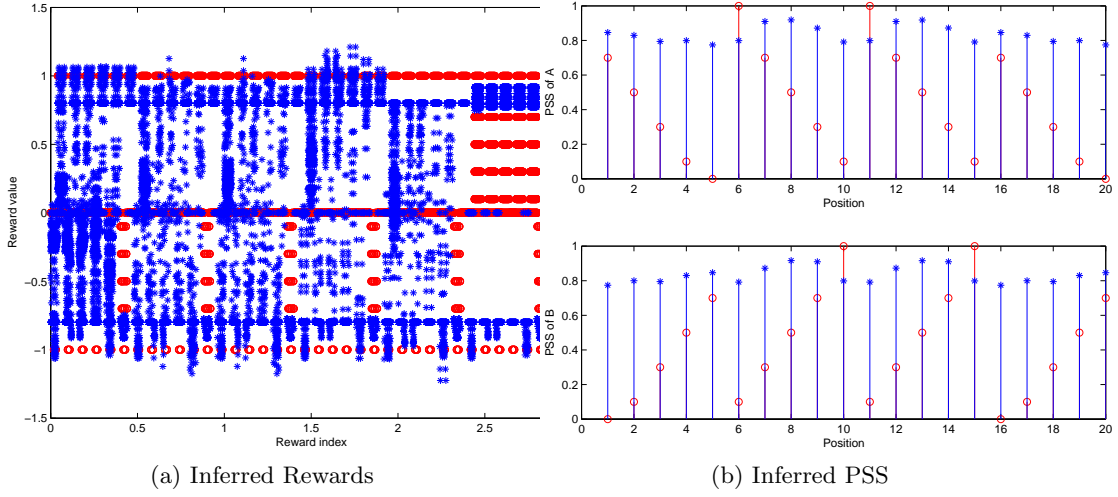


Figure 3: Inferred rewards and PSS: weak mean & weak covariance

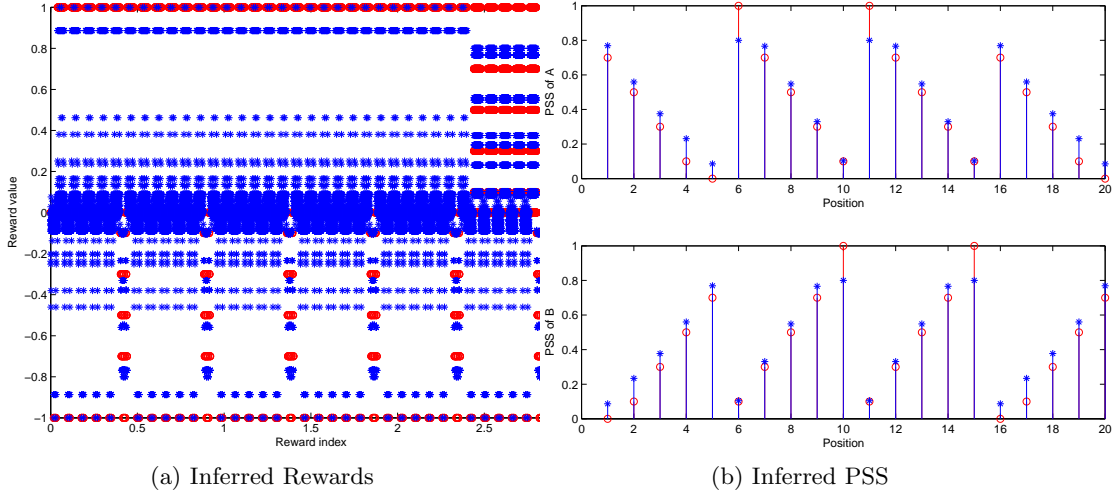


Figure 4: Inferred rewards and PSS: weak mean & strong covariance

#### 4.5 Analysis of Results

From Figures 3-8 and Table 2, we can come to the following conclusions:

- The closer the mean is to the actual rewards, the better the quality of learned rewards will be, and likewise for the covariance matrix.
- The covariance matrix has a greater influence on the quality of learned than does the mean.

From Figure 9 we can see that our MIRL algorithm successfully learns the goals for A and B.

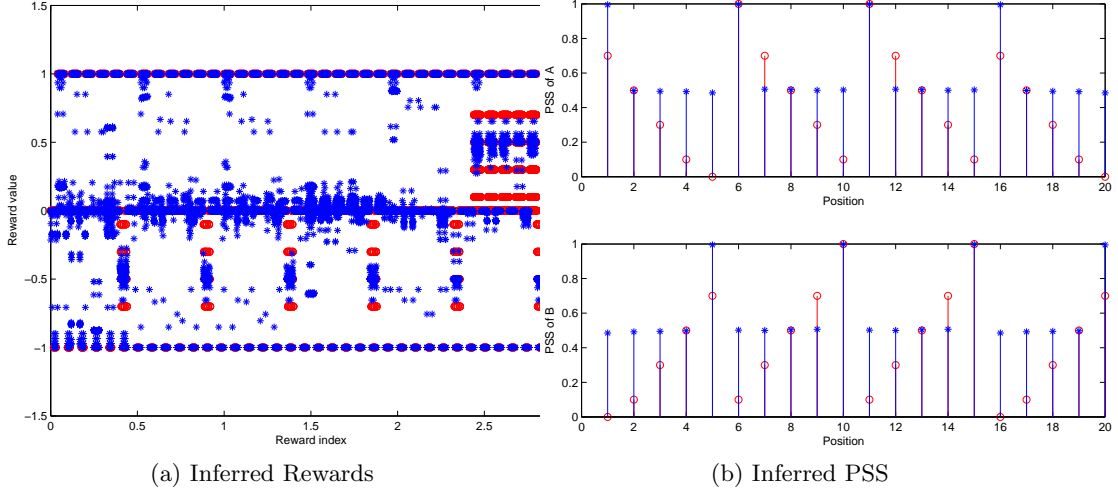


Figure 5: Inferred rewards and PSS: median mean & weak covariance

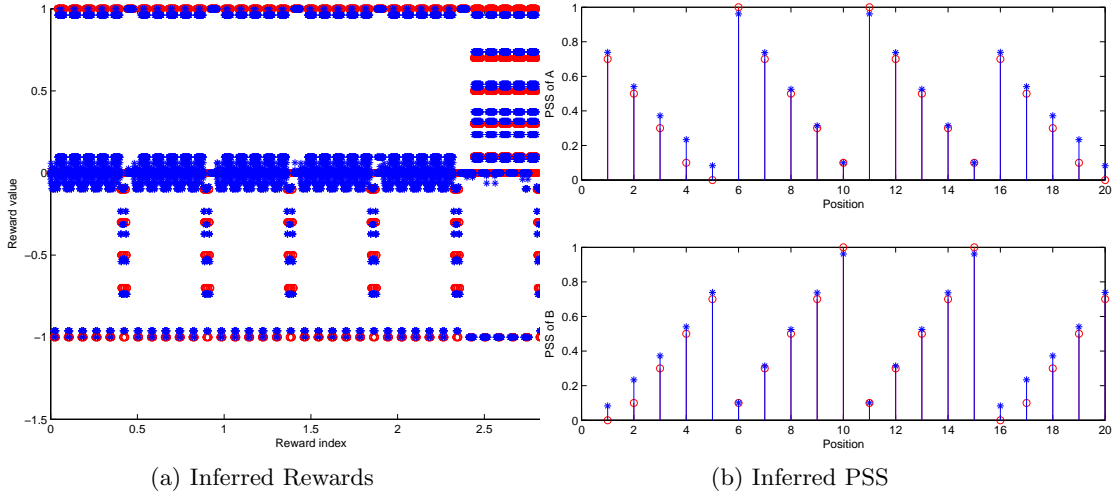


Figure 6: Inferred rewards and PSS: median mean & strong covariance

Finally, Figure 10 shows that the smaller the  $\beta$  is, the less accurate the recovered PSS will be. The rationale behind it is that players are inclined to dribble the ball rather than shoot it toward their opponents' goal when  $\beta$  is smaller, and consequently, observing the strategy of dribbling will not generate constraints that substantially alter the mode of the priors on rewards associated with shooting. For example, when  $\beta = 0.2$ , the probability of successfully dribbling the ball to the destination for each player is, at worst,  $(1 - \beta)^4 = 0.407$ , which means that a shot will never be taken in positions where the agent's PSS is 0.3 or 0.1.

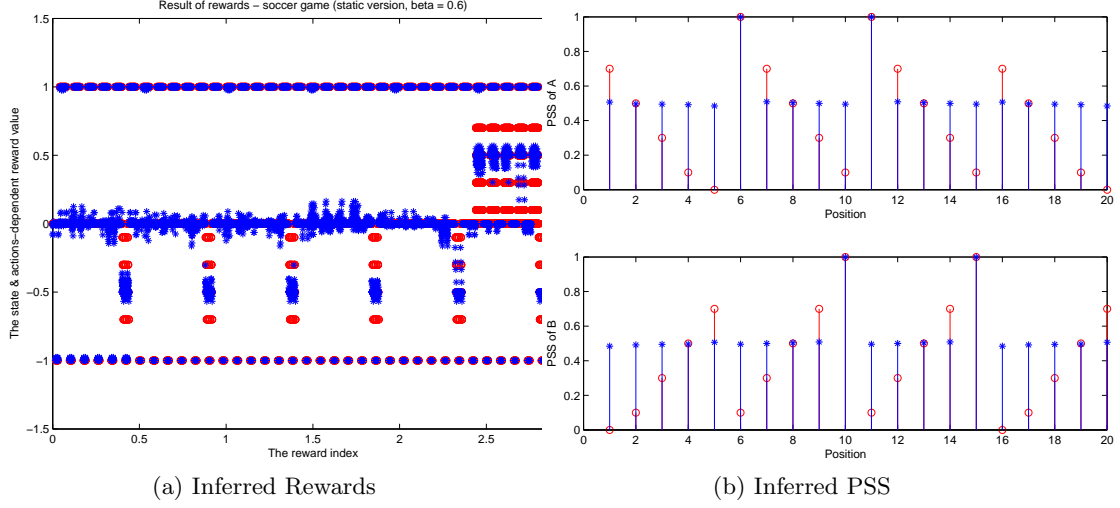


Figure 7: Inferred rewards and PSS: strong mean & weak covariance

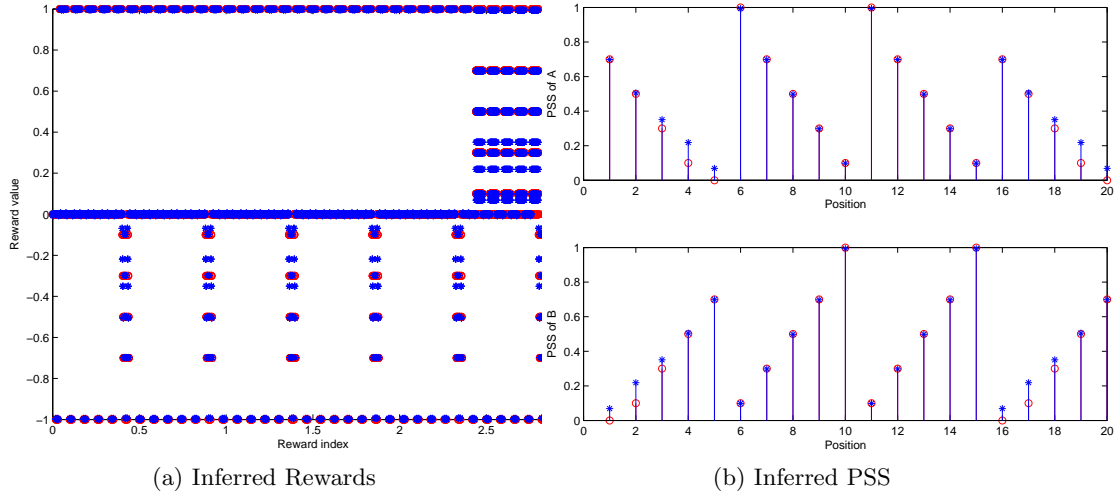


Figure 8: Inferred rewards and PSS: strong mean & strong covariance

## 5. Monte Carlo Simulation using Recovered Rewards

In the previous section distance metrics in reward and PSS space are used to evaluate the quality of learned rewards. In this section we measure reward quality in terms of the quality of the forward solution that would be based on the rewards. IRL is often set in the context of apprenticeship learning, in which learned rewards form the basis for anticipating or mimicking the response of agents to unknown situations. In MIRL, the analogous notion is to use learned rewards as the basis for game play in different environmental settings.

Consider the soccer example. We discussed and presented results of a typical case where  $\beta$ , the ball exchange rate, equals 0.6. Suppose this game between A and B is observed by a third agent C, who wants to play with B in other situations, such as  $\beta = 0.4$ . Obviously,

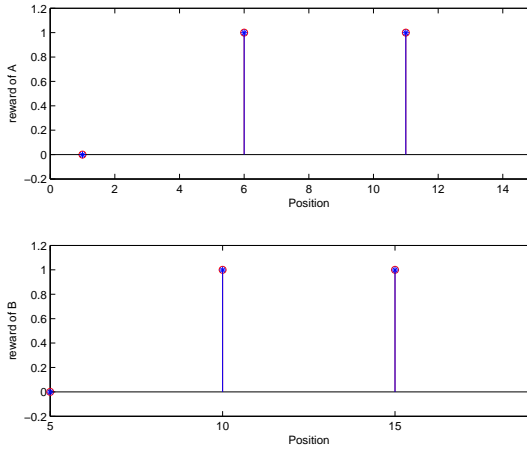


Figure 9: Actual goal recovery in case # 6

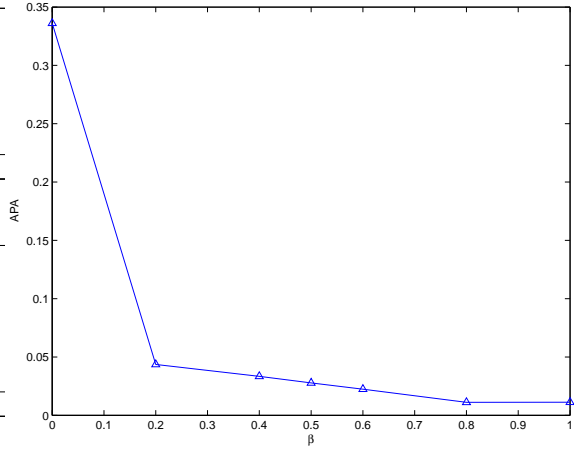


Figure 10: APD with  $\beta$  changing

B's policy in the new situation will change. But as long as C has learned A's rewards (and hence B's rewards as well because of the zero-sum property), she could develop a minimax policy, which is optimal for her, to play against B.

In this section, we will simulate games between B and C, where the ball exchange rate varies. Being rational, B will employ a minimax policy, based off of the true rewards. C will also follow a minimax policy, based off of her learned rewards. Recall that we have recovered 6 sets of rewards by assuming 6 different priors, in the previous section, we will compare the win-lose results of cases where different sets of rewards are employed.

The simulation results are presented in Table 3. In this table, the first column is the rewards that C employs to develop her minimax policy, where *true* means the true rewards and *random* denotes the rewards based on which result in a *random policy*, that is, all possible actions are taken in equal probability in every state. *WM*, *MM*, *SM*, *WC* and *SC* stand for *weak mean*, *median mean*, *strong mean*, *weak covariance matrix* and *strong covariance matrix*, respectively. The rest columns are the simulation results of 10000 rounds of games between B and C in cases where  $\beta$  being 0.4, 1 and 0. For a more clear comparison, we only count those game episodes ending in win-lose outcomes. In each column, the first and second number are C's *win* and *loss* percentages, respectively.

Let us coin the term *Application Metric* (AM) to refer to C's probability of winning in the soccer example. We can draw two conclusions from Table 3. First, B outperforms or ties C in general. That conclusion is reasonable because B knows the true rewards so that his policy is truly optimal. The second conclusion comes from Figure 11, which compares AM with the previous numerical metric ARD. As expected, a larger ARD results in a smaller probability of winning. What is notable is the sudden crash in probability of winning experienced when ARD becomes sufficiently large. Equivalently, the probability of win drops sharply when both the mean and covariance are weak. The implication is that inferring the structure of the unknowns, is much more crucial than inferring their true values.

| Base Rewards | C vs B ( $\beta = 0.4$ ) | C vs B ( $\beta = 1$ ) | C vs B ( $\beta = 0$ ) |
|--------------|--------------------------|------------------------|------------------------|
| True         | 49.95% vs 50.05%         | 50.08% vs 49.92%       | 49.43% vs 50.57%       |
| Random       | 22.78% vs 77.22%         | 19.22% vs 80.78%       | 22.91% vs 77.09%       |
| WM & WC      | 0.00% vs 100.00%         | 0.00% vs 100.00%       | 0.00% vs 100.00%       |
| WM & SC      | 49.58% vs 50.42%         | 49.78% vs 50.22%       | 50.66% vs 49.43%       |
| MM & WC      | 37.62% vs 62.38%         | 38.95% vs 61.05%       | 36.29% vs 63.71%       |
| MM & SC      | 49.70% vs 50.30%         | 49.63% vs 50.37%       | 49.84% vs 50.16%       |
| SM & WC      | 37.63% vs 62.37%         | 37.32% vs 62.68%       | 49.31% vs 50.69%       |
| SM & SC      | 49.16% vs 50.84%         | 49.75% vs 50.25%       | 49.84% vs 50.06%       |

Table 3: C vs B games simulation results

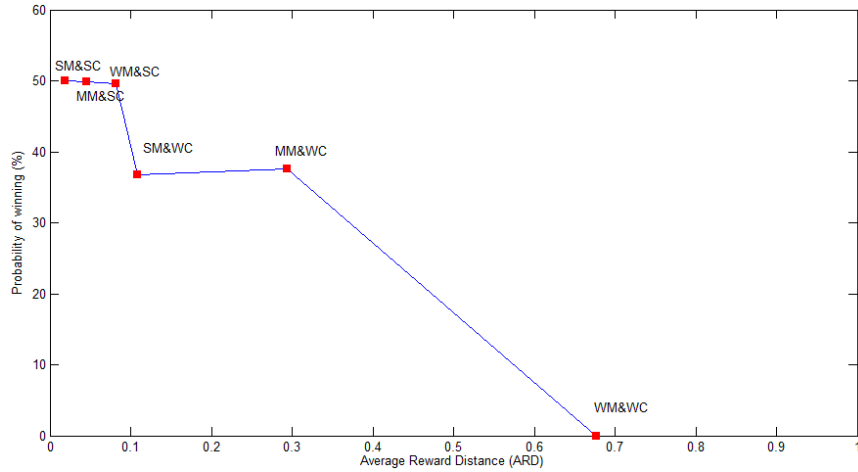


Figure 11: Two metrics comparison

## 6. Conclusions and Future Research

In this paper we introduced the MRL problem in the setting of zero-sum stochastic games and presented a solution based on Bayesian inference. Although it seems that MRL is a natural extension of IRL, the MRL problem presents new challenges. The soccer example in this paper demonstrates the difficulties we may encounter in a two-person zero-sum MRL problem in the real world. Because of the ill-specified nature of the MRL problem, it may be difficult to obtain good inverse learning results without knowing anything other than observations of policies. Fortunately in many real problems, additional information that can help structure the prior may be available.

Even in simple static games, two important distinctions between inverse learning for optimization and inverse learning for games emerge. While the model taken in this paper assumes that the complete bipolicy of two players is observed, it is more likely that only actions of the individual players are observed. In an optimization setting, since deterministic policies are assumed, strategies can be inferred exactly from finitely many observations of actions. In the case of games, strategies are often mixed, and so strategies cannot be inferred exactly from finitely many observations of the actions taken in each state. Therefore, we



cannot model a player’s strategy as an observation as can be done in IRL. In the setting of games, strategies must be treated as latent variables that are not observed directly, but bridge the gap between reward functions and observable actions.

Another direction in the future is likely to be the discovery of appropriate generative models for more general problems, including perhaps the two-person general-sum game. General-sum games engender unique challenges that complicate the development of likelihood functions. The primary challenges associated with games result from the non-uniqueness of equilibria and the possibly stochastic nature of equilibrium strategies. In general games, multiple equilibria may be associated with different game values for the two players. Specifying a likelihood function requires assuming an equilibrium selection mechanism for the two players. For example, we might assume that players choose a strategy uniformly at random from among all available equilibria associated with the reward functions. As another alternative, we might assume that players are driven toward equilibria that generate a greater value for that player. The specific assumptions imposed on equilibrium selection will affect the nature of the reward functions recovered from an inverse learning procedure.

## Acknowledgements

This work is partially supported by Science Applications International Corporation (SAIC) through the Research Scholars Fellowships Program.

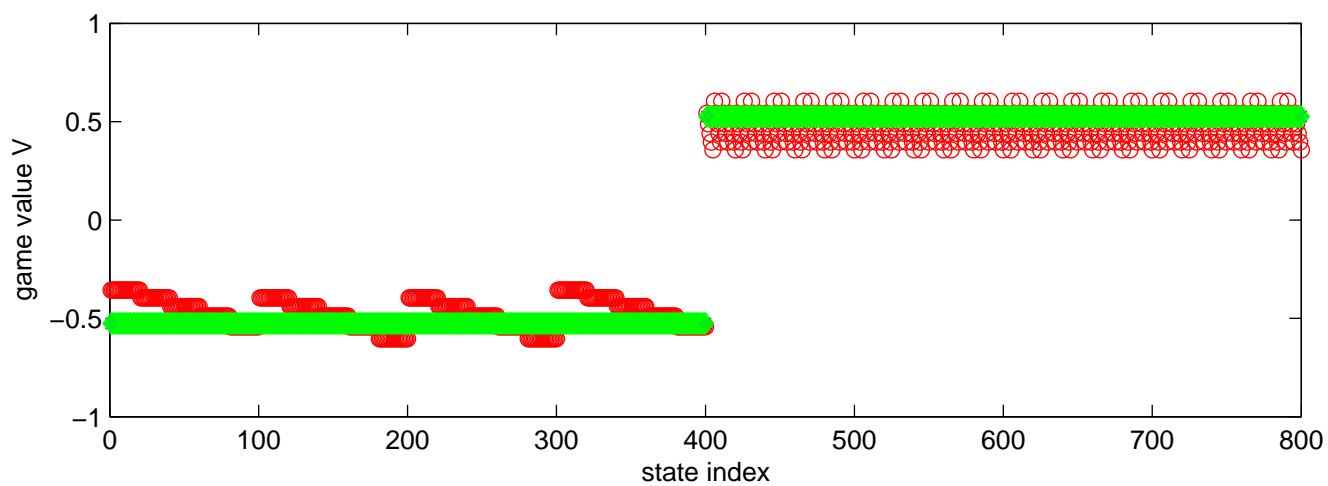
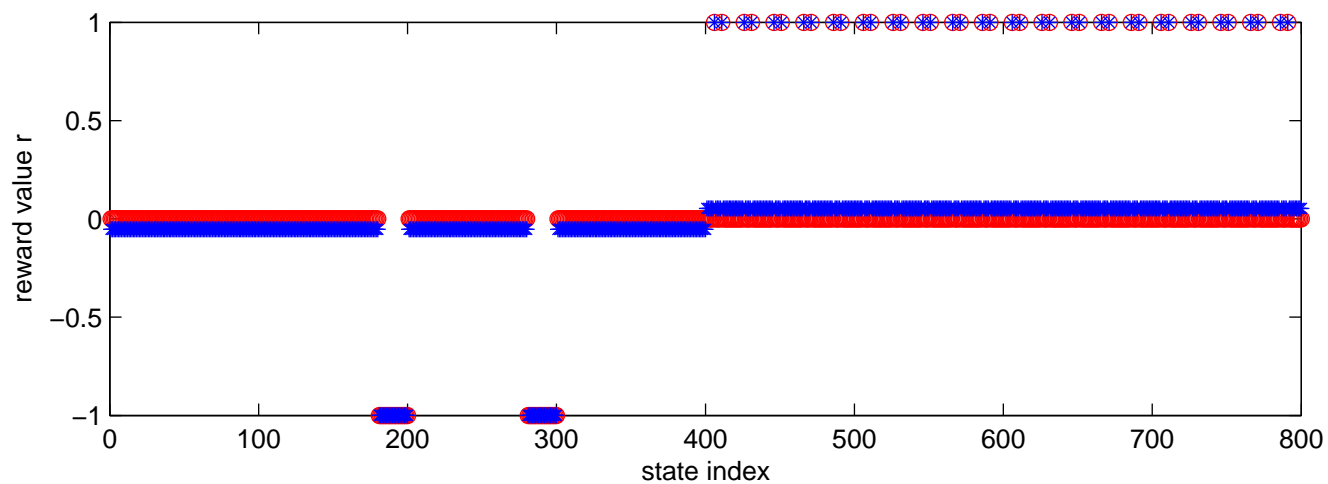
## References

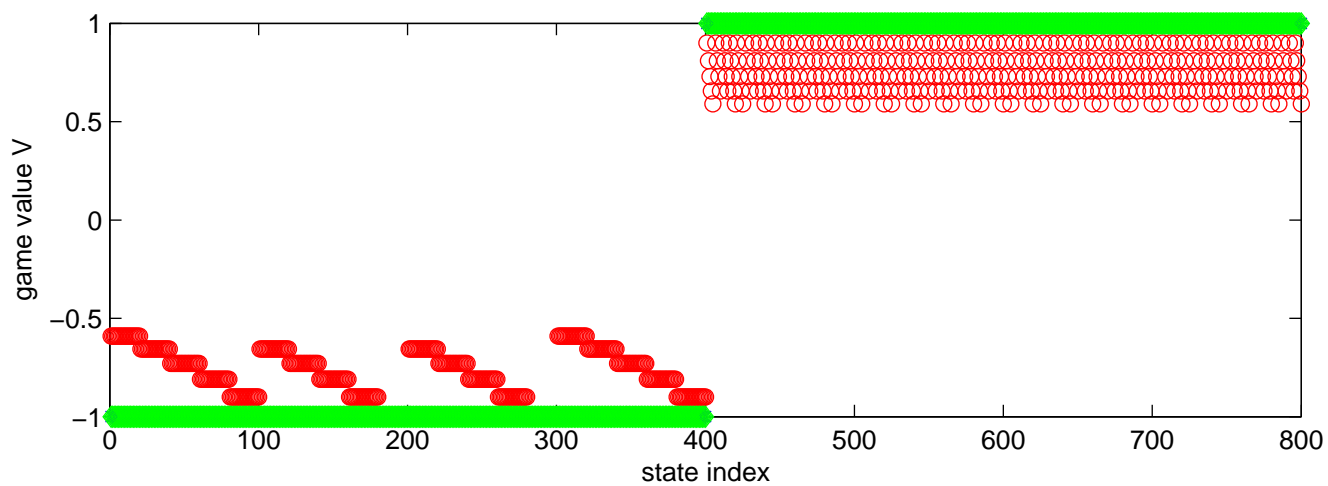
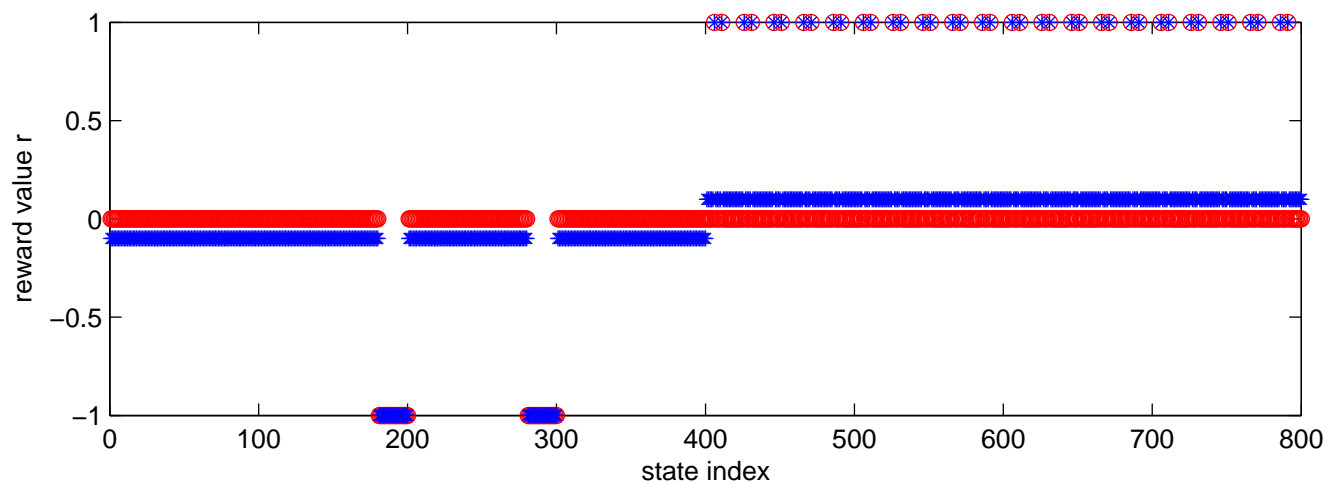
- Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21th International Conference on Machine Learning, ICML’04*, pp. 1–8.
- Abdallah, S., & Lesser, V. (2008). A multiagent reinforcement learning algorithm with non-linear dynamics. *Journal of Artificial Intelligence Research*, 33, 521–549.
- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3), 329–349.
- Beling, P., Cogill, R., & Lin, X. (2013). Multiagent inverse reinforcement learning for zero-sum stochastic games. In *51st Annual Allerton Conference on Communication, Control and Computing*.
- Bertsekas, D. P. (2005). *Dynamic Programming and Optimal Control* (3rd edition)., Vol. 1. Athena Scientific, Belmont, MA.
- Bewley, T., & Kohlberg, E. (1976). The asymptotic theory of stochastic games. *Mathematics of Operations Research*, 1(3), 197–208.
- Choi, J., & Kim, K. (2011). Map inference for bayesian inverse reinforcement learning. In *Proceedings of the 24th Advances in Neural Information Processing Systems, NIPS’01*, pp. 1989–1997.

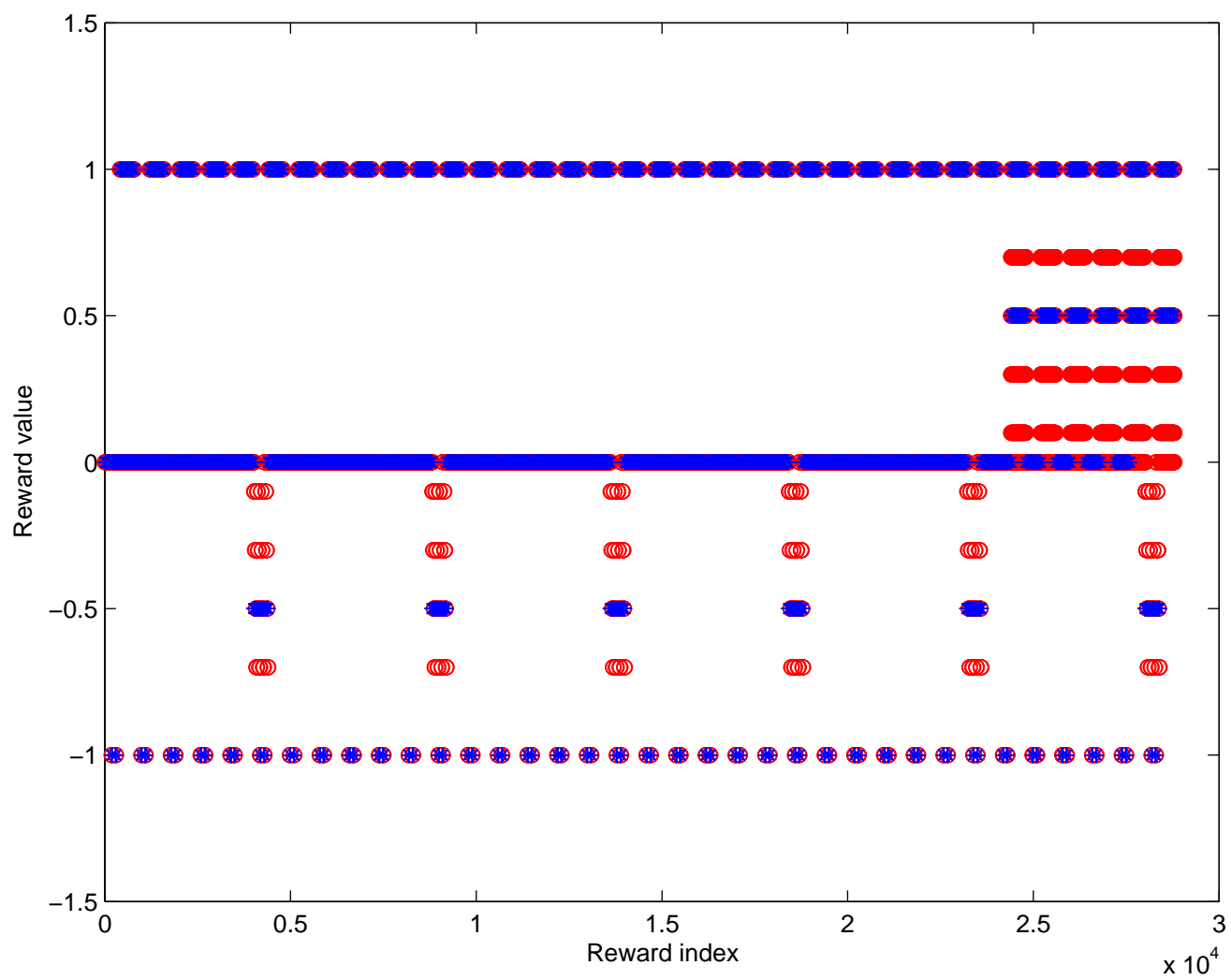
- Dimitrakakis, C., & Rothkopf, C. A. (2011). Bayesian multitask inverse reinforcement learning. In *Proceedings of the 9th European Workshops on Reinforcement Learning, EWRL'11*, pp. 273–284.
- Engel, Y., Mannor, S., & Meir, R. (2005). Reinforcement learning with gaussian processes. In *Proceedings of the 22nd International conference on Machine learning, ICML '05*, pp. 201–208.
- Federgruen, A. (1980). Successive approximation methods in undiscounted stochastic games. *Operations Research*, 28(3), 794–809.
- Ferguson, T. S. (2008). *Game Theory*. UCLA.
- Filar, J., & Vrieze, K. (1996). *Competitive Markov Decision Processes* (1st edition). Springer-Verlag, New York, NY.
- Higham, N. (2002). *Accuracy and Stability of Numerical Algorithms* (2nd edition). SIAM.
- Hu, J., & Wellman, M. P. (1998). Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the 15th International Conference on Machine Learning, ICML'98*, pp. 242–250.
- Kash, I., Friedman, E., & Halpern, J. (2011). Multiagent learning in large anonymous games. *Journal of Artificial Intelligence Research*, 40, 571–598.
- Krishnamurthy, D., & Todorov, E. (2010). Inverse optimal control with linearly-solvable mdps. In *Proceedings of the 27th International Conference on Machine Learning, ICML'10*, pp. 335–342.
- Kushner, H. J., & Chamberlain, S. G. (1969). Finite state stochastic games: Existence theorems and computational procedures. *IEEE Transactions on Automatic Control*, AC-14(3), 248–255.
- Lee, S. J., & Popovic, Z. (2010). Learning behavior styles with inverse reinforcement learning. *ACM Transactions on Graphics*, 29(4), 122:1–122:7.
- Levine, S., Popović, Z., & Koltun, V. (2011). Nonlinear inverse reinforcement learning with gaussian processes. In *Proceedings of the 24th Advances in Neural Information Processing, NIPS'11*, pp. 19–27.
- Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning, ICML'94*, pp. 157–163.
- Maitra, A., & Parthasarathy, T. (1970). On stochastic games. *Journal on Optimization Theory and Applications*, 5(4), 289–300.
- Mertens, J. F., & Neyman, A. (1980). Stochastic games. *International Journal of Game Theory*, 10(2), 53–66.
- Michini, B., & How, J. P. (2012). Bayesian nonparametric inverse reinforcement learning. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD'12*, Vol. 2, pp. 148–163.
- Monash, C. A. (1979). *Stochastic Games: The Minimax Theorem*. Ph.D. thesis, Department of Mathematics, Harvard University, Cambridge, MA.

- Natarajan, S., Kunapuli, G., Judah, K., Tadepalli, P., Kersting, K., & Shavlik, J. W. (2010). Multi-agent inverse reinforcement learning. In *Proceedings of the 9th International Conference on Machine Learning and Applications, ICMLA'10*, pp. 395–400.
- Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.
- Ng, A. Y., & Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning, ICML'00*, pp. 663–670.
- Owen, G. (1968). *Game Theory* (1st edition). W. B. Saunders Company, Philadelphia, PA.
- Patek, S. D., Beling, P. A., & Zhao, Y. (2007). Natural solutions for a class of symmetric games. In *AAAI Spring Symposium: Game Theoretic and Decision Theoretic Agents*, pp. 47–53.
- Qiao, Q., & Beling, P. A. (2011). Inverse reinforcement learning via convex programming. In *Proceedings of the 2011 American Control Conference, ACC'11*, pp. 113–118.
- Raghavan, T. E. S., & Filar, J. A. (1991). Algorithms for stochastic games - a survey. *Methods and Models of Operations Research (Zeitschrift für O. R.)*, 35(6), 437–472.
- Ramachandran, D., & Amir, E. (2007). Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI'07*, pp. 2586–2591.
- Rao, S. S., Chandrasekaran, R., & Nair, K. P. K. (1973). Algorithms for discounted stochastic games. *Journal of Optimization Theory and Applications*, 11(6), 627–637.
- Rogers, P. D. (1969). *Nonzero-Sum Stochastic Games*. Ph.D. thesis, Engineering Science, Graduate Division, University of California, Berkeley, CA.
- Russell, S. (1998). Learning agents for uncertain environments (extended abstract). In *Proceedings of the 11th Annual Conference on Computational Learning Theory, COLT'98*, pp. 101–103.
- Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences, Mathematics*, 39, 1095–1100.
- Sobel, M. J. (1971). Noncooperative stochastic games. *The Annals of Mathematical Statistics*, 42(6), 1930–1935.
- Vrieze, O. J. (1987). *Stochastic Games with Finite State and Action Spaces*. Centrum voor Wiskunde en Informatica (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands.
- Waugh, K., Ziebart, B., & Bagnell, J. (2011). Computational rationalization: The inverse equilibrium problem. In *Proceedings of the 28th International Conference on Machine Learning, ICML'11*, pp. 1169–1176.
- Yang, E., & Gu, D. (2004). Multiagent reinforcement learning for multi-robot systems: A survey. Technical report CSM-404, University of Essex.
- Zhao, Y., Patek, S., & Beling, P. (2008). Decentralized bayesian search using approximate dynamic programming methods. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 38(4), 970–975.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference of Artificial Intelligence, AAAI'08*, Vol. 3, pp. 1433–1438.







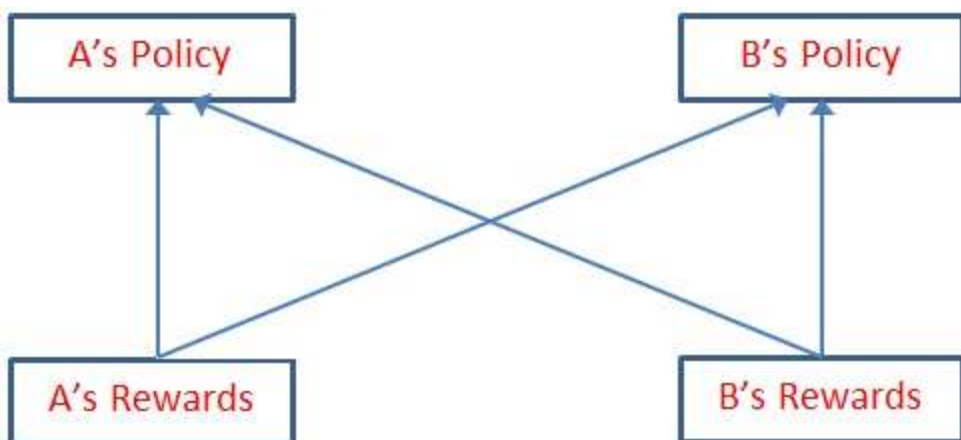
This figure "AB-relationship2.jpg" is available in "jpg" format from:

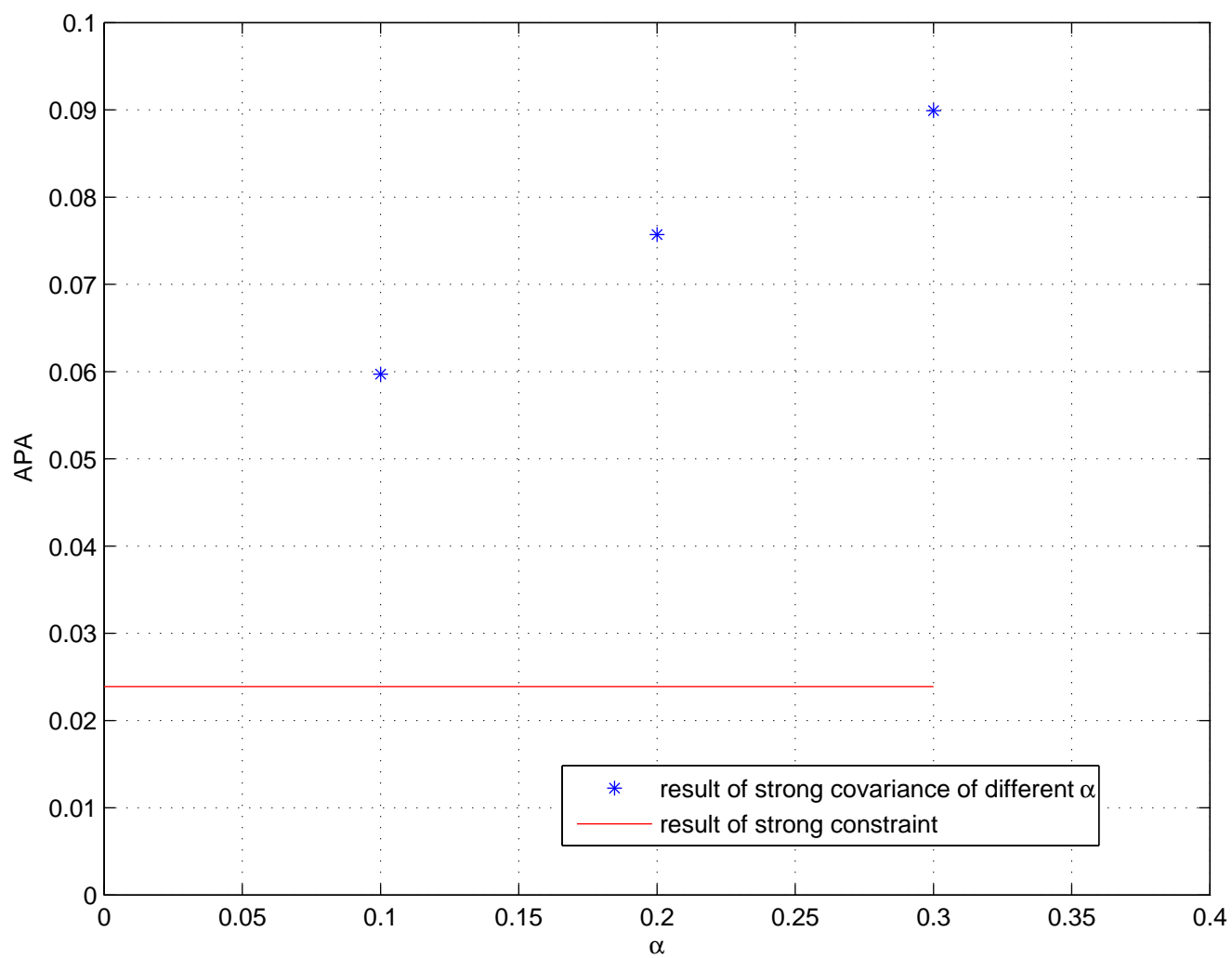
<http://arxiv.org/ps/1403.6508v1>



This figure "AB\_relationship.jpg" is available in "jpg" format from:

<http://arxiv.org/ps/1403.6508v1>





This figure "fancy\_two\_matrices\_comparison.png" is available in "png" format from:

<http://arxiv.org/ps/1403.6508v1>

